Decoding MT Motion Response for Optical Flow Estimation: An Experimental Evaluation

N V Kartheek Medathati, Manuela Chessa, Guillaume S. Masson, Pierre Kornprobst, Fabio Solari

Decoding MT Motion Response for Optical Flow Estimation: An Experimental Evaluation

N. V. Kartheek Medathati, Manuela Chessa, Guillaume S. Masson, Pierre Kornprobst, Fabio Solari

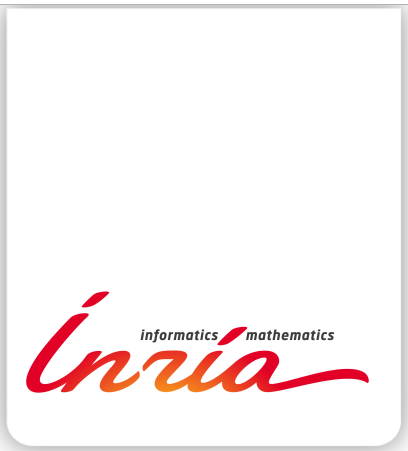
RESEARCH REPORT

N° 8696

March 2015

Project-Team Neuromathcomp

ISSN 0249-6399 ISRN INRIA/RR--8696--FR+ENG



Decoding MT Motion Response for Optical Flow Estimation: An Experimental Evaluation

N. V. Kartheek Medathati*†, Manuela Chessa‡, Guillaume S. Masson§, Pierre Kornprobst†, Fabio Solari§

Project-Team Neuromathcomp

Research Report n° 8696 — March 2015 — 11 pages

Abstract: Motion processing in primates is an intensely studied problem in visual neurosciences and after more than two decades of research, representation of motion in terms of motion energies computed by V1-MT feedforward interactions remains a strong hypothesis. Thus, decoding the motion energies is of natural interest for developing biologically inspired computer vision algorithms for dense optical flow estimation. Here, we address this problem by evaluating four strategies for motion decoding: intersection of constraints, maximum likelihood, linear regression on MT responses and neural network based regression using multi scale-features. We characterize the performances and the current limitations of the different strategies, in terms of recovering dense flow estimation using Middlebury benchmark dataset widely used in computer vision, and we highlight key aspects for future developments.

Key-words: Bio-inspired approach, optical flow, spatio-temporal filters, motion energy, population code, V1, MT, Middlebury dataset

* Both authors N. V. Kartheek Medathati and Manuela Chessa should be considered as first author.
† INRIA, Neuromathcomp team, Sophia Antipolis, France
‡ University of Genova, DIBRIS, Italy
§ Institut des Neurosciences de la Timone, CNRS, Marseille, France

RESEARCH CENTRE
SOPHIA ANTIPOLIS – MÉDITERRANÉE

2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex

Décoder les Réponses de MT pour Estimer le Flot Optique: Une Evaluation Expérimentale

Résumé : Le traitement du mouvement chez les primates est un problème largement étudié en neurosciences et après plus de deux décennies de recherches, la représentation du mouvement en terme d'énergies de mouvement obtenues par une intégration montante entre les aires corticales V1-MT reste une hypothèse forte. Ainsi, le décodage de ces énergies de mouvement est d'un intérêt naturel pour développer des algorithmes de vision par ordinateur bio-inspiré pour l'estimation d'un flot optique dense. Ici, nous abordons ce problème en évaluant quatre stratégies pour le décodage du mouvement: Intersection de contraintes, maximum de vraisemblance, régression linéaire sur les réponses de MT et méthode de regression basée sur un réseau de neurone utilisant l'ensemble des échelles. Nous caractérisons les performances et les limites actuelles des différentes stratégies, en termes d'estimation de flot optique denses en utilisant la base de test Middlebury faisant référence en vision par ordinateur, et nous mettons en évidence quelques idées clés pour les développements futurs.

Mots-clés : Approche bio-inspirée, flot optique, filtres spatio-temporels, énergie de mouvement, codage en population, V1, MT, base de test Middlebury

Contents

1 Introduction

Visual motion estimation is a widely studied problem in both computer vision and Visual Neuroscience. How do primates estimate motion? has been a question of intense focus in visual neuroscience yet only partly understood owing both to underlying complexity and to the experimental stimuli that has been used, see [1] for a recent review. The limitations of the experimental and modeling studies in motion estimation so far have been well explained by Nishimoto et al. [2], in terms of partial coverage in spatio-temporal frequency domain, e.g., only direction of motion [3, 4] or two-dimensional slice [5, 6]. Though in [2] the authors show that the widely accepted feed-forward spatio-temporal filtering model is a good fit for explaining neural responses to naturalistic videos, the model has not been tested in terms of recovering the dense velocity vector field, called optical flow, which has been extensively studied in computer vision due to its broad application potential.

Recovering dense optical flow shows a limitation of these models based on spatio-temporal filters, how spatial acuity of the motion is preserved and how does the model deal with several naturalistic scenarios such as motion boundaries, occlusions, and transparencies. Modern computer vision datasets with ground truth, such as Middlebury dataset [7], give us an opportunity to study these aspects also with respect to the problem of decoding. The goal of this paper is to evaluate four decoding strategies to estimate optical flow from an MT motion response.

This paper is organised as follows. In Sect. 2, we present the basis of this approach which is a feedforward model of V1 and MT cortical areas response. It is a summary of the model presented in [8] in which we revisited the seminal work by Heeger and Simoncelli [9, 10]: Model includes V1 simple and complex cells to estimate motion energy [11] and MT pattern cells [4, 10]. In Sect. 3, given an MT motion response, we propose four decoding strategies to estimate optical flow. These four strategies are then evaluated and discussed in Sect. 4 using classical sequences from the literature.

2 V1-MT model for motion processing

2.1 Area V1: Motion Energy

Let us consider a grayscale image sequence $I(p,t)$, for all positions $p = (x, y)$ inside a domain Ω and for all time $t > 0$. Our goal is to find the optical flow $v(p,t) = (v_x, v_y)(p,t)$ defined as the apparent motion at each position p and time t.

Simple cells are characterized by the preferred spatial orientation θ of their contrast sensitivity in the spatial domain and their preferred velocity v^c in the direction orthogonal to their contrast orientation often referred to as component speed. The receptive fields of the V1 simple cells are classically modeled using band-pass filters in the spatio-temporal domain. In order to achieve low computational complexity, the spatio-temporal filters are decomposed into separable filters in space and time. Spatial component of the filter is described by Gabor filters h and temporal component by an exponential decay function k. We define the following complex filters:

$$\begin{aligned} h(p;\theta,f_s) =& Be^{\left(\frac{-(x^2+y^2)}{2\sigma^2}\right)} e^{j2\pi(f_s cos(\theta)x + f_s sin(\theta)y)}, \\ k(t;f_t) =& e^{\left(-\frac{t}{\tau}\right)} e^{j2\pi(f_t t)}, \end{aligned}$$

where σ and τ are the spatial and temporal scales respectively, which are related to the spatial and temporal frequencies f_s and f_t and to the bandwidth of the filter. Denoting the real and imaginary components of the complex filters h and k as h_e, k_e and h_o, k_o respectively, and a

preferred velocity (speed magnitude) $v_c = f_t/f_s$, we introduce the odd and even spatio-temporal filters defined as follows,

$$\begin{aligned} g_o(p,t;\theta,v^c,\sigma) =& h_o(p;\theta,f_s)k_e(t;f_t) + h_e(p;\theta,f_s)k_o(t;f_t), \\ g_e(p,t;\theta,v^c,\sigma) =& h_e(p;\theta,f_s)k_e(t;f_t) - h_o(p;\theta,f_s)k_o(t;f_t). \end{aligned}$$

These odd and even symmetric and tilted (in space-time domain) filters characterize V1 simple cells. Using these expressions, we define the response of simple cells, either odd or even, with a preferred direction of contrast sensitivity θ in the spatial domain, with a preferred velocity v^c and with a spatial scale σ by

$$R_{o/e}(p,t;\theta,v^c,\sigma) = g_{o/e}(p,t;\theta,v^c,\sigma) \overset{(p,t)}{*} I(p,t). \tag{1}$$

The complex cells are described as a combination of the quadrature pair of simple cells (1) by using the motion energy formulation,

$$E(p,t;\theta,v^c,\sigma) = R_o(p,t;\theta,v^c,\sigma)^2 + R_e(p,t;\theta,v^c,\sigma)^2,$$

followed by a normalization. Assuming that we consider a finite set of orientations $\theta = \theta_1 \dots \theta_N$, the final V1 response is given by

$$E^{V1}(p,t;\theta,v^c,\sigma) = \frac{E(p,t;\theta,v^c,\sigma)}{\sum_{i=1}^{N} E(p,t;\theta_i,v^c,\sigma) + \varepsilon}, \tag{2}$$

where $0 < \varepsilon \ll 1$ is a small constant to avoid divisions by zero in regions with no energies, which happens when no spatio-temporal texture is present.

2.2 Area MT: Pattern Cells Response

MT neurons exhibit velocity tuning irrespective of the contrast orientation. This is believed to be achieved by pooling afferent V1 responses in both spatial and orientation domains followed by a non-linearity [10]. The response of a MT pattern cell tuned to the speed v^c and to direction of speed d can be expressed as follows:

$$E^{MT}(p,t;d,v^c,\sigma) = F\left(\sum_{i=1}^{N} w_d(\theta_i)\mathcal{P}(E^{V1})(p,t;\theta_i,v^c,\sigma)\right),$$

where w_d represents the MT linear weights that give origin to the MT tuning (see example in Fig. 1). It can be defined by a cosine function shifted over various orientations [4, 12], i.e.,

$$w_d(\theta) = \cos(d-\theta) \quad d \in [0, 2\pi[.$$

Then, $\mathcal{P}(E^{V1})$ corresponds to the spatial pooling and is defined by

$$\mathcal{P}(E^{V1})(p,t;\theta_i,v^c,\sigma) = \frac{1}{A}\sum_{p'} f_\alpha(\|p-p'\|)E^{V1}(p,t;\theta_i,v^c,\sigma), \tag{3}$$

where $f_\alpha(s) = \exp(s^2/2\alpha^2)$, $\|.\|$ is the L_2-norm, α is a constant, A is a normalization term (here equal to $2\pi\alpha^2$) and $F(s) = exp(s)$ is a static nonlinearity chosen as an exponential function [4]. The pooling defined by (3) is a spatial Gaussian pooling.

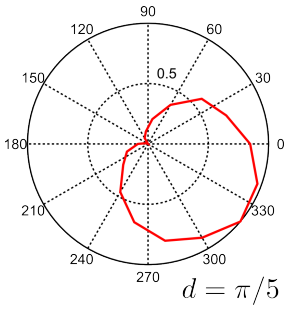


Figure 1: Example of a MT direction ($d = \pi/5$) tuning curve for moving plaid stimuli that span all the speed directions.

3 Decoding of the velocity representation of area MT

In order to engineer an algorithm capable of recovering dense optical flow estimates, we need to address the problem of decoding the population responses of tuned MT neurons. Indeed, a unique velocity vector cannot be recovered from the activity of a single velocity tuned MT neuron as multiple scenarios could evoke the same activity. However, a unique vector can be recovered from the population activity. In this paper, the velocity space was sampled by considering MT neurons that span over the 2-D velocity space with a preferred set of tuning speed directions d in $[0, 2\pi[$ and also a multiplicity of tuning speeds v^c.

Four strategies are described below. The first three strategies, called *intersection of constraints*, *maximum likelihood* and *learned linear* decoding are based on coarse-to-fine approach in order to consider multiple spatial frequencies f_s and to compute large velocities. This approach is illustrated in Figure 2 and described in [8]. Here the decoding stage will impact the quality of the optical flow extracted at each scale and used for the warping. In the third strategy an optimal linear decoding is learned and is applied at each scales. Alternatively, the fourth strategy, called *regression with neural network*, learns to estimate optical flow from the V1 responses at every scales.

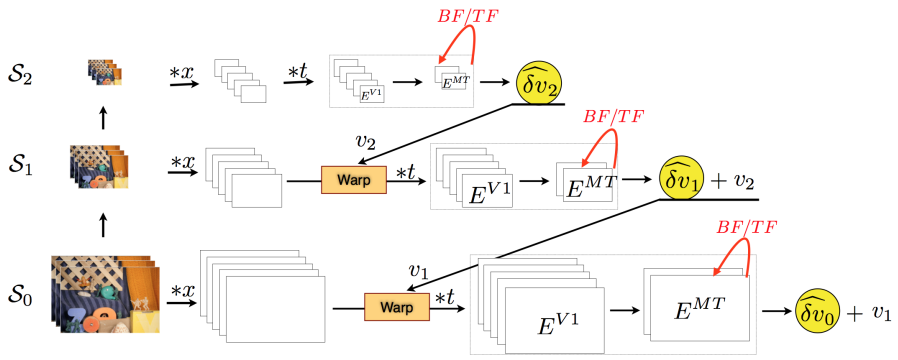


Figure 2: Coarse-to-fine approach for optical flow based on a V1-MT model [8]. At each scale, decoding is needed to warp V1 motion energies at the coarser scale.

3.1 Intersection of Constraints Decoding

The MT responses are obtained through a static nonlinearity described by an exponential function, thus we can linearly decode the population activities [13]. Since the distributed representation of velocity is described as a function of two parameters (speed and direction), first we linearly decode the speed (velocity magnitude) for each speed direction, then we apply the intersection of constraints (IOC) mechanism [1] to compute the speed direction. The speed along direction d can be expressed as:

$$v^d(p,t;d,\sigma) = \sum_{v_i^c=v_1^c}^{v_M^c} v_i^c E^{MT}(p,t;d,v_i^c,\sigma). \tag{4}$$

Then the IOC solution is defined by:

$$\begin{array}{l} \vec{v} = \underset{\vec{w}}{\operatorname{argmin}} \{G(\vec{w})\}, \\ \text{where } G(\vec{v}) = \sum_{d_i=d_1}^{d_Q} (v^{d_i} - \vec{v} \cdot [\cos d_i \sin d_i]^T)^2, \end{array} \tag{5}$$

where $(\cdot)^T$ indicates the transpose operation. The analytic solution of Eq. 5 gives:

$$\begin{array}{rcl} v_x & = & \frac{2}{Q} \sum_{d_i=d_1}^{d_Q} v^d(p,t;d_i,\sigma) \cos d_i \\ v_y & = & \frac{2}{Q} \sum_{d_i=d_1}^{d_Q} v^d(p,t;d_i,\sigma) \sin d_i. \end{array} \tag{6}$$

3.2 Maximum Likelihood Decoding

The MT activities (see Fig 3 for an example of a MT population response that shows a peak for the direction and the speed present in the input stimulus) can be decoded with a Maximum Likelihood (ML) technique [14]. In this paper, the ML estimate is performed through a curve

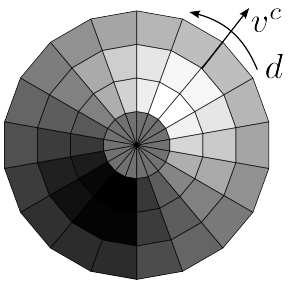


Figure 3: An example of MT population response at a given image point p, for a random dot sequence that moves at $v_x = 0.3$ and $v_y = 0.3$ pixel/frame. The speed directions d have 19 values in the range $[0, 2\pi[$, and the speeds v^c have 7 values in the range ± 1 pixel/frame.

fitting, or template matching, method. In particular, we decode the MT activities by finding the Gaussian function that best match the population response. The position of the peak of the Gaussian corresponds to the ML estimate.

3.3 Linear Decoding Through Learned Weights

We can learn the two-dimensional matrix of weights $\mathcal{W}$ that are used to linearly decode the MT activities ($\vec{v} = E^{MT}\mathcal{W}$ for each image pixel p). To learn such weights, we have considered a

dataset of 8×7 random dot sequences with known speeds (both v_x and v_y, 8 directions and 7 speeds), which cover the spatio-temporal filters' range, and we have minimized a cost function to compute the best weights $\mathcal{W}$. The cost function is defined by:

$$||\mathcal{R}\mathcal{W} - v_{gt}||^2 + \lambda||\mathcal{W}||^2, \tag{7}$$

where $\mathcal{R}$ is a matrix whose rows contain the MT population responses (for the whole training set), $\mathcal{W}$ is the vector of weights, and v_{gt} contains the ground truth speeds. It is worth to note that such procedure has been carried out at a single spatial scale. Since we use random dots, we have considered the average MT responses, and $\lambda = 0.05$. Figure 4 shows the learned two-dimension matrix of weights.

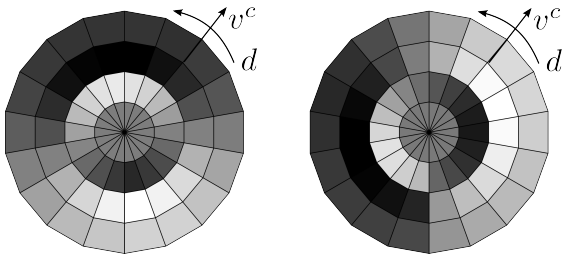


Figure 4: Two-dimensional matrix of weights learned through sequences of random dots. The matrices on the left and on the right are used to decode v_x and v_y, respectively.

3.4 Decoding with Regression using Neural Network

For the regression using neural network, spatio-temporal energies representative of the V1 complex cell responses are computed across various scales and are concatenated to form an input vector of dimension 504 (6 scales $\times$ 12 orientations $\times$ 7 velocities). The feature computation stage is illustrated in Fig. 5. It is worth to note that in this decoding strategy we do not use the coarse to fine approach. A feedforward network comprising of a hidden sigmoidal layer and a linear output layer with 400 neurons in the hidden layer and 2 neurons in the output layer, computing velocity along x and y axis is considered. The hidden layer can be interpreted as MT cells tuned to different velocities. For training the network, subsampled features by a factor of 30 from Middlebury sequences are used and the network is trained for 500 epochs using back propagation algorithm till the RMSE of the network over the training samples has reached 0.3. Note that we only have a single network or a regressor and it is applied to all pixels. For training and simulating the experiment PyBrain package has been used.

4 Experimental Evaluation and Discussion

Table 1 shows the average angular errors (AAE) and the end-point errors (EPE), and the corresponding standard deviations, by considering the Middlebury training set and the Yosemite sequence. Results for the four decoding strategies (intersection of constraints, maximum likelihood, linear decoding with learned weights, and regression using Neural Networks) are reported. Some sample optical flows for the four decoding methods are reported in Figure 6. The results show that the intersection of constraints approach gives estimates similar to the ones obtained by considering a linear decoding through learned MT weights. A fitting with Gaussian functions to implement a maximum likelihood decoding does not perform as well as the IOC approach:

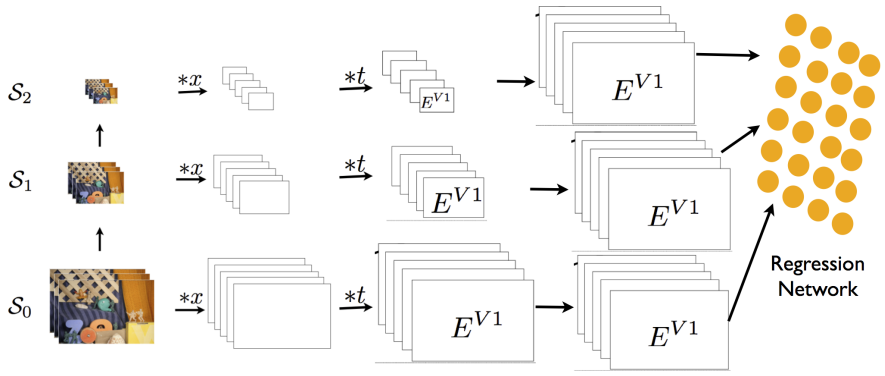


Figure 5: Scale space for regression based learning.

this is due to the actual MT activity pattern, and to the fact that MT population responses for low speed has several peaks and it is hard to fit a Gaussian.

Observing the results obtained after decoding suggests that scale-space with warping procedure is not well suited for analysis with spatio-temporal features and is inducing larger errors when compared to the regression scheme where the spatio-temporal motion energies across scales are simultaneously taken into consideration. This is in accordance with earlier model by Heeger, where plane fitting in spatio-temporal domain has been adapted, indicating that interscale interactions are critical in velocity decoding. The neural network based regression has preserved motion edges much better when compared to the warping scheme in most of the sequences, but however it fails in the Yosemite sequence, which indicates that there is some diffusion happening in regions without motion energy as could be seen in the sky region. The responses of the network need to be more smooth to better match the ground truth, however this is to be expected as this regression scheme does not have any neighborhood interactions and smoothness criterion in place. This needs to be further investigated by incorporating spatial pooling of the motion energies and spatial interactions at the MT level into the model. On the whole, this indicates that restoring spatial acuity of motion estimation by population decoding is a little studied problem and there is a large scope for improvement as the current decoding schemes do not perform on par with state of the art results in computer vision.

	Intersection of constraints		Maximum Likelihood		Learned Weights		Regression using NN	
Sequence	AAE ± STD	EPE ± STD	AAE ± STD	EPE ± STD	AAE ± STD	EPE ± STD	AAE ± STD	EPE ± STD
grove2	**4.33** ± 10.28	**0.30** ± 0.62	9.78 ± 21.08	0.74 ± 1.30	4.59 ± 9.69	0.32 ± 0.59	5.17 ± 8.49	0.37 ± 0.54
grove3	**9.65** ± 19.02	1.14 ± 1.83	13.73 ± 25.70	1.47 ± 2.32	9.94 ± 18.79	1.15 ± 1.79	9.67 ± 15.39	**1.01** ± 1.42
Hydrangea	5.98 ± 11.19	0.62 ± 0.97	8.88 ± 20.41	0.85 ± 1.44	6.34 ± 11.83	0.65 ± 1.00	**3.22** ± 6.21	**0.29** ± 0.41
RubberWhale	10.16 ± 17.73	0.34 ± 0.54	16.28 ± 26.31	0.73 ± 1.45	10.07 ± 16.65	0.34 ± 0.51	**7.61** ± 8.98	**0.25** ± 0.26
urban2	5.21 ± 10.17	0.58 ± 1.06	14.24 ± 20.37	1.51 ± 1.94	16.46 ± 22.81	1.49 ± 1.91	**4.59** ± 9.69	**0.32** ± 0.59
urban3	15.78 ± 35.94	1.90 ± 3.24	18.24 ± 39.45	1.82 ± 2.91	14.05 ± 33.29	1.74 ± 3.07	**5.76** ± 17.49	**0.80** ± 1.51
Yosemite	**3.49** ± 2.86	**0.16** ± 0.16	5.34 ± 7.24	0.31 ± 0.69	3.80 ± 2.98	0.18 ± 0.18	20.09 ± 14.74	0.86 ± 0.87
all	9.14 ± 16.86	0.85 ± 1.35	12.36 ± 22.94	1.06 ± 1.72	9.32 ± 16.56	0.84 ± 1.29	**8.02** ± 11.57	**0.56** ± 0.80

Table 1: Error measurements on Middlebury training set and on the Yosemite sequence.

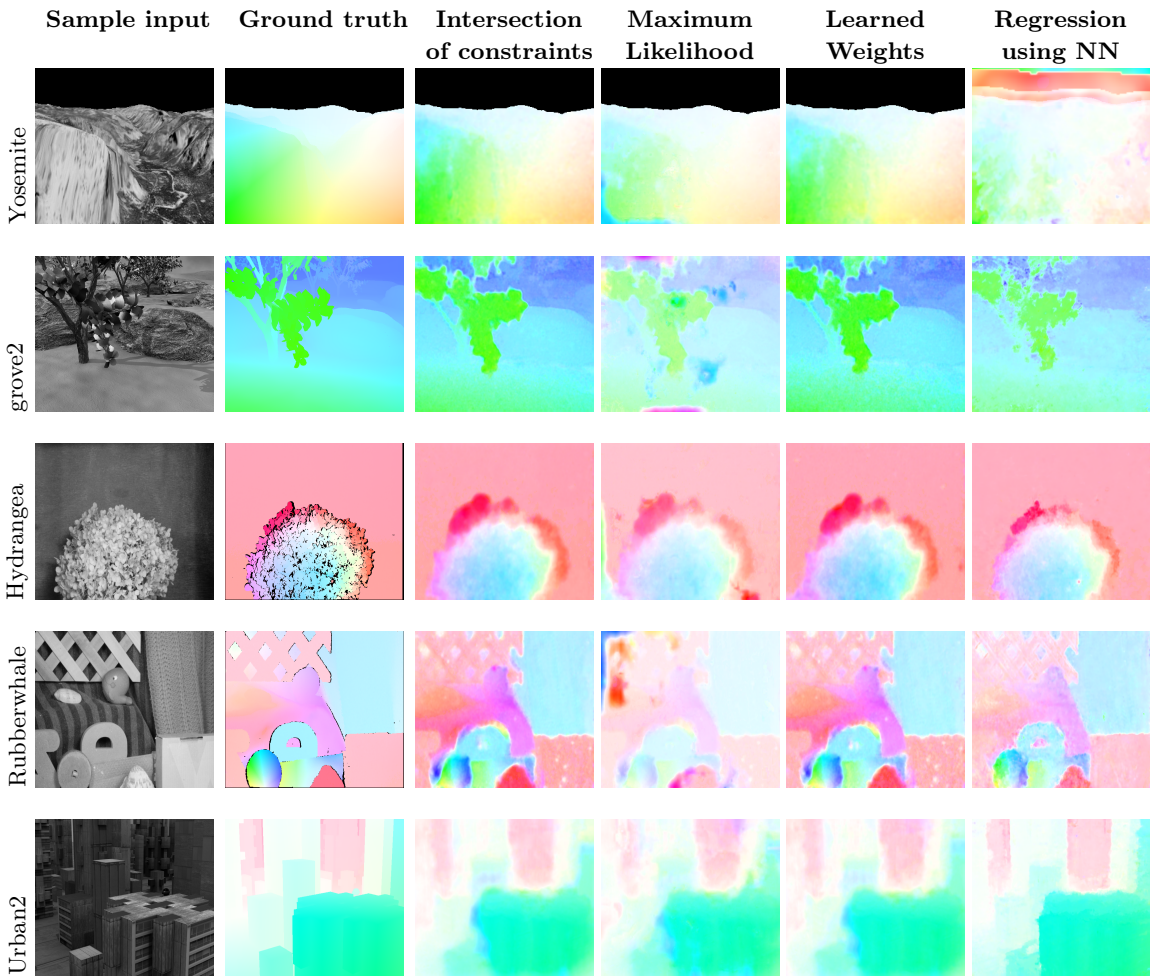


Figure 6: Sample results on a subset of Middlebury training set and on the Yosemite sequence.

Acknowledgments

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 318723 (MATHEMACS) and grant agreement no. 269921 (BrainScaleS).

References

[1] D. C. Bradley and M. S. Goyal, "Velocity computation in the primate visual system," *Nature Reviews Neuroscience*, vol. 9, no. 9, pp. 686–695, 2008. 4, 7

[2] S. Nishimoto and J. L. Gallant, "A three-dimensional spatiotemporal receptive field model explains responses of area mt neurons to naturalistic movies," *The Journal of Neuroscience*, 2011. 4

[3] C.C. Pack and R.T. Born, "Temporal dynamics of a neural solution to the aperture problem in visual area MT of macaque brain," *Nature*, 2001. 4

[4] N C Rust, V Mante, E P Simoncelli, and J A Movshon, "How MT cells analyze the motion of visual patterns," *Nature Neuroscience*, vol. 9, no. 11, pp. 1421–1431, 2006. 4, 5

[5] J. Perrone and A. Thiele, "Speed skills: measuring the visual speed analyzing properties of primate mt neurons," *Nature Neuroscience*, 2001. 4

[6] N. Priebe, C. Cassanello, and S. Lisberger, "The neural representation of speed in macaque area MT/V5," *Journal of Neuroscience*, 2003. 4

[7] S. Baker, D. Scharstein, J.P. Lewis, S. Roth, M. Black, and R. Szeliski, "A database and evaluation methodology for optical flow," *International Journal of Computer Vision*, vol. 92, no. 1, pp. 1–31, 2011. 4

[8] F. Solari, M. Chessa, K. Medathati, and P. Kornprobst, "What can we expect from a classical V1-MT feedforward architecture for optical flow estimation?," Research Report. Accepted to Image Communication RR-8618, 2014. 4, 6

[9] D.J. Heeger, "Optical flow using spatiotemporal filters," vol. 1, no. 4, pp. 279–302, Jan. 1988. 4

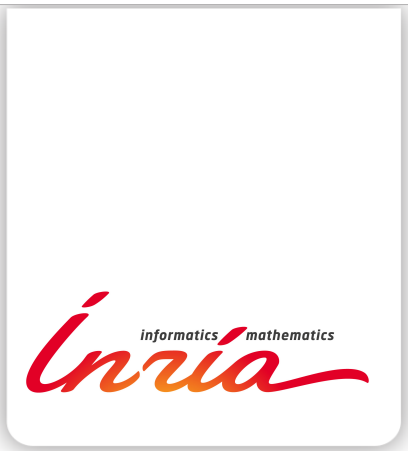
[10] E. P. Simoncelli and D. J. Heeger, "A model of neuronal responses in visual area MT," *Vision Research*, vol. 38, no. 5, pp. 743 – 761, 1998. 4, 5

[11] E.H. Adelson and J.R. Bergen, "Spatiotemporal energy models for the perception of motion," *J. of the Opt. Soc. of America A*, vol. 2, pp. 284–299, 1985. 4

[12] J. H. Maunsell and D. C. Van Essen, "Functional properties of neurons in middle temporal visual area of the macaque monkey. I. selectivity for stimulus direction, speed, and orientation," *Journal of Neurophysiology*, vol. 49, no. 5, pp. 1127–1147, 1983. 5

[13] K. R. Rad and L. Paninski, "Information rates and optimal decoding in large neural populations.," in *NIPS*, 2011, pp. 846–854. 7

[14] A. Pouget, K. Zhang, S. Deneve, and P. E. Latham, "Statistically efficient estimation using population coding," *Neural Comp.*, vol. 10, no. 2, pp. 373–401, 1998. 7



RESEARCH CENTRE
SOPHIA ANTIPOLIS – MÉDITERRANÉE

2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399